**Descriptions of individual differences measures of L1 skills used in various sites**

**Dutch**

***Boston Naming Test*** The short Dutch version of the Boston Naming Test (‘Verkorte Boston BenoemTaak’; Roomer, Brok, Hoogerwerf, & Linn, 2011), a standardized and validated picture naming task. To ease the completion of the task, a computerized version of this task was created in E-Prime 2 (Psychology Software Tools, Pittsburgh, PA). Participants saw 29 pictures and typed their answer. Scoring was performed in accordance with the instructions of the Dutch version and corrected for age and education level (Roomer et al., 2011). The reliability for this task was α = .61; test duration is 5-10min.

Roomer, E. K., Brok, S., Hoogerwerf, A. C., & Linn, D. E. (2011). *Boston BenoemTaak. Een test voor woordvinding.*

***LexTALE – Dutch*** The Dutch version of the LexTALE vocabulary test (Lemhöfer & Broersma, 2012), which is a standardized and validated lexical decision task with 40 word trials and 20 nonword trials. The LexTALE was presented on a display and programmed in Tscope (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006); participants had to press the appropriate keys to indicate whether the stimulus was a word or a nonword. Scoring was performed in accordance with the instructions of Lemhöfer and Broersma (2012). The reliability for this task was α = .61; test duration is 3-5min.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: A C library for programming cognitive experiments on the MS windows platform. *Behavior Research Methods*, *38*(2), 280–286.

***GL&SCHR*** A Dutch spelling test in which participants have to write 30 dictated words on paper (De Pessemier & Andries, 2009). A pre-recorded audio file is used to dictate the word list to ensure a fixed timing, participants have to write the words while the list is playing. They have the opportunity to repeat the word list. One correctly spelled word equals one point. The reliability for this task was α = .60; test duration is 4-5min.

De Pessemier, P., & Andries, C. (2009). *GL&SCHR*. Leuven/Apeldoorn: Garant.

**English**

***Sight word efficiency and Phonemic decoding efficiency*** subtests from Form A of the Test of Word Reading Efficiency: TOWRE-2 (Torgesen et al., 2012). Materials and instructions are available in Dropbox/multi-lingual eye tracking reading database/tests/offline/TestsEnglish. The score is a number of correctly pronounced items (words or non-words) within 45 seconds.

Torgesen, J. K., Wagner, R., & Rashotte, C. (2012). *Test of Word Reading Efficiency:(TOWRE-2)*. Pearson Clinical Assessment.

***Vocabulary size test*** adapted from Nation and Beglar’s (2007). The 14,000 word version of the original test of receptive vocabulary contains 140 multiple-choice items, with 10 items from each 1000 word family level. Each item contains a target word and a sentence in which this word is used (sometimes in a different morphological form). Four definitions for the target word follow, numbered a-d. The instruction is to “Circle the letter a-d with the closest meaning to the key word in the question”.

Example:

SEE

They **saw** it.

1. cut
2. waited for
3. looked at
4. started

This original test was adapted to enable more rapid assessment. The first ten items (representing 1000 most common words of English) were skipped, and all participants started from the 2000 word family level. Also, a stop rule was established such that a subject responded to ten items representing the same word family level (i.e., 1000 words) and was only allowed to move to the next word family level (next 1000) if they committed less than 5 errors in those 10 items. The test stopped if a subject had more than 5 errors in a given 1000 word family level or completed the test. The score is the number of correct responses.

https://www.wgtn.ac.nz/lals/about/staff/Publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I.S.P. &Beglar, D. (2007) A vocabulary size test. *The Language Teacher, 31(7)*, 9-13.

***Spelling recognition test*** adopted from Andrews and Hersch (2010). The original test consists of 88 English words, of which half is spelled correctly and half is not. Given a very high split-half reliability of this original test among McMaster students (r = 0.88; unpublished data) and in consideration of time, we only used odd items from the original list: out the 44 items, 22 were spelled correctly and 22 were not. The score is the number of correct responses.

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, *139*(2), 299.

**Subject motivation** self-report measure, 10-item Student Opinion Survey (SOS; Thelk, Sundre, Horst, & Finney, 2009).

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. The Journal of General Education, 58, 129-151.

See also: <http://www.rpajournal.com/dev/wp-content/uploads/2016/07/A1_Corrected.pdf>

***The English vocabulary test LexTALE*** (Lemhöfer & Broersma, 2012) a lexical decision test that consists of 60 items (40 words, 20 nonwords). The score is the percentage of correct responses, corrected for the unequal proportion of words and nonwords in the test by averaging the percentages correct for these two item types (http://www.lextale.com/scoring.html).

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*(2), 325-343.

**Estonian**

***Estonian Author Recognition Test (ART).*** The test is modeled after Stanovich et al. (1989) and measures individual differences in exposure to print in Estonian. Participants are provided with a list containing 100 names out of which 50 are Estonian authors and 50 are Estonian persons who are not authors. The task of the participants is to mark all authors with a cross. Authors represented different genres and were chosen from the list of best-selling authors in Estonia in the years 2015-2020 according to Estonian bookstore bestseller lists of various categories. The non-authors were created by randomly combining common Estonian first and last names taken from Statistics Estonia’s name statistics. It was checked that the created names were actually not well-known person names. The author list contained 19 female names and 31 male names, the non-author list contained 26 female names and 24 male names. 1 point is given for each correct answer, 0 for a blank answer, and -1 for an incorrect answer. The maximum score is thus 50. This test has no time limit, but participants are instructed to answer promptly rather than spending too long thinking about each name. The test was administered in a pen and paper format.

Stanovich, K. E., &; West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*, 402–433.

***Estonian Verbal Fluency task (L1 vocabulary tests).***In this test participants are instructed to name aloud as many words as possible for a specific category within 60 seconds. In the current study participants were asked to name words from two different semantic categories and two different phonemic categories. Semantic category fluency was tested with animals and clothes and phonemic fluency with words starting with ”k” and “p”. Participants were instructed that morphological variants (e.g., dog; dogs) and proper nouns do not count as valid words. After transcription, the amount of valid words was counted for each category. Harrison et al. (2000) presented a test–retest reliability results of r = 0.82 for phonological fluency and r = 0.68 for semantic fluency. The test was administered on a computer using Powerpoint. Responses were recorded and transcribed afterwards.

Harrison JE, Buxton P, Husain M, Wise R. (2000) Short test of semantic and phonological fluency: normal performance, validity and test–retest reliability. *British Journal of Clinical Psychology, 39*, 181–91.

***Estonian Big Five Personality Inventory (Big5).*** The personality of the participants was assessed on five dimensions (openness, conscientiousness, extraversion, agreeableness, neuroticism) using an Estonian translation of the 44 item Big Five inventory of Personality (John, O. P., & Srivastava, S. 1999). The test was administered on a computer using Google Forms.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York: Guilford Press.

**Finnish**

***Finnish Author Recognition Test*** The test is modeled after Stanovich et al. (1989) and measures individual differences in exposure to print in Finnish. Participants are provided with a list containing 100 names out of which 50 are Finnish authors and 50 are Finnish persons who are not authors. The task of the participants is to mark all authors with a cross. Authors represented different genres and were chosen from the list of best-selling authors in Finland in the year 2016-2017. The non-authors were taken from the editorial boards of two medical magazines. For both the authors and non-author group an equal amount of men and women were selected. 1 point is given for each correct answer, 0 for a blank answer, and -1 for an incorrect answer. The maximum score is thus 50, the minimum score in principle -50. This test has no time limit, but participants are instructed to answer promptly rather than spending too long time thinking about each name.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly,* 24*,* 402–433.

***Finnish Verbal Fluency task*** In this test participants are instructed to name aloud as many words as possible for a specific category within 60 seconds. In the current study participants were asked to name words from two different semantic categories and two different phonemic categories. Semantic category fluency was tested with *animals* and *clothes* and phonemic fluency with words starting with ”k” and “p”. Participants were instructed that morphological variants (dog => dogs) and proper nouns do not count as valid words. The test was recorded with a dictation machine. After transcription, the amount of valid words was counted for each category. Harrison et al. (2000) presented test–retest reliability results of r = 0.82 for phonological fluency and r = 0.68 for semantic fluency.For the current study we calculated the amount of correct words for the two phonemic categories together, for the two semantic scores together, and the overall score of all four categories taken together.

Harrison JE, Buxton P, Husain M, Wise R. (2000) Short test of semantic and phonological fluency: normal performance, validity and test–retest reliability. *British Journal of Clinical Psychology*, *39*, 181–91.

***Lexize – Vocabulary test*** Lexize is a Finnish version of Lextale (see Brysbaert, 2013; Izura et al., 2014; Lemhöfer & Broersma, 2012) developed by Salmela, Lehtonen, Garusi and Bertram. The test is a simple lexical decision task in which participants have to indicate whether a given letter string is a Finnish word or nonword. In the current test 132 items (88 words and 44 nonwords) were included with words ranging from very low to high frequency. The test was administered in a computerized form with Experiment Builder software. The test is designed to assess vocabulary knowledge of L2 speakers from A1 to C2 level in the Common European Framework of Reference for Languages (CEFR). When performed by L2 speakers of Finnish, instructions emphasize that there is no time limit to answer. As Finnish university students typically perform at ceiling level in Lexize, the current experiment’s instructions emphasized that items should be responded to as fast and accurately as possible. Hence we also report RT as dependent variable. For these, error responses and RTs longer than 2000 ms have been excluded. Error rates are reported in two different ways. First the raw number of errors is reported; second the LexizeScore is reported which is similar to how scores are calculated in LexTale tests (with a larger penalty for responding incorrectly to nonwords).

This LexizeScore is calculated as follows:

LexizeScore = N CorrectlyRespondedWords – 2\*N(IncorrectlyResponded Nonwords). The maximum score in this experiment is thus 88.

The test can be performed via <https://psyk.abo.fi/LexizeWeb/#/> or downloaded from all app stores (see also <https://psyk.abo.fi/Lexize/Lexize.html>).

Brysbaert, M. (2013). LEXTALE\_FR: A fast, free, and efficient test to measure language

proficiency in French. *Psychologica Belgica, 53,* 23-37.

# Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. Psicologica, 35, 49-66.

Lemhöfer, K., Broersma, M. (2012) Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* 44, 325–343.

**German**

***Vocabulary knowledge***Students were administered form A of the CFT 20 Vocabulary Knowledge Test (Weiß, 2006), which comprises 30 multiple-choice items. On each trial, one word was given (e.g., “chaos”) and participants had to select a word with the same or a similar meaning from a set of five different words (e.g., “[a] confusion, [b] damage, [c] currency, [d] anger, or [e] demonstration”). The test encompassed 30 trials of increasing difficulty over trials and participants were given 5 minutes to complete as many as possible. The number of correctly solved trials served as the score.

Weiß, R. H. (2006). *CFT 20-R.* *Grundintelligenztest Skala 2 (CFT 20-R) [Basic Intelligence Scale 2 (CFT 20-R)]*. Göttingen, Germany: Hogrefe.

***Reading fluency*** Word and nonword reading were tested using the one-minute reading tests of the Salzburger Lese- und Rechtschreibtest (SLRT-II; Moll & Landerl, 2010). In this standardized reading test, 156 words are presented over eight columns and the participant's task is to accurately read out loud as many words as possible in one minute. The nonword part is exactly the same only that it contains pronounceable nonwords instead of words. The number of correctly read words/nonwords then gives the achieved test score.

Moll, K., & Landerl, K. (2010). *Lese- und Rechtschreibtest (SLRT II)*. Bern, Switzerland: Huber.

**Greek**

***Reading fluency***.Participants’ word reading skills were assessed with two measures, a real word and a nonword reading task, adapted in Greek from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999; Greek adaptation: Kendeou, Papadopoulos, & Spanoudis, 2015).

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efﬁciency*. Austin, TX: PRO-ED.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2015). Reading comprehension and PASS theory. In T. C. Papadopoulos, R. K. Parrila, & J. R. Kirby (Eds.), *Cognition, intelligence, and achievement* (pp. 117-136). San Diego, CA: Academic Press.

*Word Reading Fluency (WRF)* In this task, participants were given a list of 88 words and asked to read them as quickly as possible. The words were ordered in terms of difficulty, based on frequency (high/low), orthographic regularity (regular/exception), and length (2-5 syllables). The words included mainly nouns with a few adjectives and verbs. A short 8-word practice list preceded the test items. The participants’ score was the number of correctly read words within a 45-s time limit. Cronbach’s alpha reliability coefficient in our sample was .83.

*Phonemic Decoding Fluency (PDF)* In this task, an 8-item practice list presented first, followed by a list of 63 nonwords, ordered in terms of difficulty, in three columns. Nonwords were derived from real words after changing two or three letters. The task started with one-syllable words and ended with five-syllabic words with the majority being two- and three-syllabic words. Participants were asked to read the nonwords as quickly as possible. The participants’ score was the number of correctly read items within a 45-s time limit. Cronbach’s alpha reliability coefficient in our sample was .75.

***Orthographic processing*** *-* ***orthographic choice.*** This task was adapted in Greek (Papadopoulos et al., 2009) from the work of Olson and colleagues (e.g., Olson, Forsberg, Wise, & Rack, 1994). It consisted of two practice- and 45 testing-items. Each target word presented together with a homophonic distractor (e.g., /χώρος/χόρος; /horos/; place). Participants had to use their orthographic knowledge to identify the word with the correct spelling between the two options. The participants’ score was the total number of items correctly answered. Cronbach’s alpha reliability coefficient in our sample was .67.

Olson, R. K., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 243–277). Baltimore, MD: Brookes.

Papadopoulos, T. C., Georgiou, G. K., & Kendeou, P. (2009). Investigating the Double-Deficit Hypothesis in Greek: Findings from a longitudinal study. *Journal of Learning Disabilities, 42*, 528-547.

***Vocabulary.*** This task was initially used by Kendeou et al. (2015). Participants were given 15 incomplete sentences, each one followed by four word-choices. Participants were asked to read the sentence carefully and select the word that fit the blank coherently. The four words were closely similar. The participants’ score was the total number of items completed correctly. Cronbach’s alpha reliability coefficient in our sample was .79.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2015). Reading comprehension and PASS theory. In T. C. Papadopoulos, R. K. Parrila, & J. R. Kirby (Eds.), *Cognition, intelligence, and achievement* (pp. 117-136). San Diego, CA: Academic Press.

**Hebrew**

***Word and pseudo-word reading***Parallel of TOWRE. Participants are asked as many words/pseudo-words in 45 seconds. Word list includes a maximum of 104 items, and the pseudo-word list includes a maximum of 100 items. Note that words are unpointed whereas pseudo-words are pointed. Score is based on number of words/pseudo-words read correctly within the time limit (within 45 seconds in each subtest). Task developed internally by Ram Frost's lab at the Hebrew University.

***Spelling of homophones*** Participants are presented with 19 sentences each including a homophone, and are asked to choose the correct spelling of the homophone from two options. Score is number of correct answers (maximum score: 19). Test was timed with a 2-minutes maximum. Test developed by the Hadad Center, Bar-Ilan University.

Schiff, R., & Ravid D. (2009). *Spelling homophonic words test.* Hadad Center for Research in Dyslexia and Reading Disorders, Bar-Iilan University, RamatGan, Israel.

***Morphological Fluency*** In each of 10 questions, participants are presented with a Hebrew root and are asked to produce as many words derive from the root as possible in a maximum of 30 seconds (per trial). Score is number of words produced in total (summed across all 10 trials). Test developed by the Hadad Center, Bar-Ilan University.

Schiff, R., Baron, A., & Kahta, S. (2009). *Morphological fluency test.* Hadad Center for Research in Dyslexia and Reading Disorders, Bar-Iilan University, RamatGan, Israel.

***Vocabulary*** Hebrew Parallel of the Shipley Institute of Living Scale. Test includes 40 questions. Participants are presented with a target word and are asked to choose the correct synonym out of 4 options. Test was timed with a 5-minutes maximum. Score is number of correct responses out of 40. Test adapted to Hebrew by Dr. Asaf Gilboa.

**Italian**

***The LexITA*** An Italian version of the Lextale (Amenta, Badan & Brysbaert, 2019). Scoring is done according to Izura, Cuetos & Brysbaert, 2014. The task was administered in a computerized form (Inquisit software) that allowed us to extract the time for task execution and compute the Inverse Efficiency Score (Time/proportion accurate).

The following measures are reported:

- LexITA\_n\_acc LexITA: Number of accurate responses

- LexITA\_prop\_acc LexITA: Proportion of accurate responses

- LexITA\_total\_time LexITA: Time for task execution (ms)

- LexITA\_ies LexITA: Inverse efficiency score (Time/proportion accurate)

- LexITA\_index LexITA: LexITA index (computed using the formula from Izura, Cuetos & Brysbaert, 2014)

Amenta, S., Badan, L., Brysbaert, M., Lex-ITA: *Uno strumento per valutare rapidamente e affidabilmente la conoscenza lessicale per apprendenti adulti di italiano come L2*. XXV Congresso Nazionale della Sezione di Psicologia Sperimentale, September 18th-20th 2019, Milan, Italy.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica, 35*(1), 49-66.

***Syntactic screening test*** A computerized task, where subjects are required to identify, out of four possible options, the sentence which matches ('is equivalent') the meaning of a target sentence. Alternative choices could be 'Contradictory', 'Compatible' or 'Distractors'. We report accuracy, chronometric and Inverse Efficiency Score data. Normative data for this test will be soon available.

The following measures are reported:

- Syntax\_equiv Syntactic Test: Number of 'Equivalent' (correct) Responses

- Syntax\_contr Syntactic Test: Number of 'Contradictory' responses

- Syntax\_comp Syntactic Test: Number of 'Compatible' responses

- Syntax\_distr Syntactic Test: Number of 'Distractor' responses

- Syntax\_n\_acc Syntactic Test: Number of accurate responses

- Syntax\_prop\_acc Syntactic Test: Proportion of accurate responses

- Syntax\_time Syntactic Test: Mean time to complete a trial (regardless of accuracy) in ms

- Syntax\_total\_time Syntactic Test: Cumulative time to do the task (ms)

- Syntax\_ies Syntactic Test: Inverse Efficiency Score (Cumulative time/ proportion accurate responses)

**Korean**

***The Korean vocabulary task*** Developed by Lee, Seong, Choi & Lowder (2019). The version used had a modification in one question out of the 60 questions in the original. Lee et al. showed that the scores of their vocabulary task were positively correlated with those of the Author Recognition Test, which in turn were correlated with the performance in an on-line lexical decision task as well as an offline reading comprehension task. That is, higher scores in the Korean vocabulary test were reliably associated with the higher scores in the Author Recognition Test, which were in turn correlated with a better performance in the lexical decision task (i.e., faster reaction times and higher accuracy scores) and in the comprehension task (i.e., higher accuracy scores).

Lee, H., Seong, E., Choi, W., & Lowder, M. W. (2019). Development and assessment of the Korean Author Recognition Test.​ ​Quarterly Journal of Experimental Psychology, 72(7)​, 1837-1846. doi: 10.1177/1747021818814461.

**Norwegian**

***Norwegian BPVS*** This task is closely modeled on the British Picture Vocabulary Scale (BPVS-II; Dunn, Dunn, Whetton, & Burley, 1997) in terms of item properties and administration procedure. The Norwegian version has been standardized (for children 3-16 years old). It consists of 12 blocks of 12 items each. Each item is composed of one (pre-recorded) word, presented auditorily, and four pictures (displayed on the screen), one of which matches the word. The participant must choose the correct picture. We used a custom computerized version of the test (self-administered). The score is the number of items correctly responded to plus the non-administered baseline (i.e., easier blocks, presumed correct).

Reference for the creation and validation of the Norwegian version:

Lyster, S.-A.H., Horn, E., & Rygvold, A.-L. (2010). Ordforråd og ordforrådsutvikling hos norske barn og unge. Resultater fra en utprøving av British Picture Vocabulary Scale, Second Edition (BPVS II) [Vocabulary and vocabulary development in Norwegian children and adolescents. Results from a trial of the BPVS-II]. *Spesialpedagogikk, 09*, 35-43.

***Norwegian TOWRE*** Closely modeled on the TOWRE in terms of item properties and administration procedure. It has not been standardized. We used one word sheet and one pseudoword sheet. The score is the number of correctly read items per minute. Actual reading time was 45 seconds but some participants completed the sheet in less time, so in every case (including L1 & L2) we divided the number of correctly read items with the actual time taken (45 s or less).

Reference for the creation and first reported use of this test:

Furnes, B., & Samuelsson, S. (2010). Predicting Reading and Spelling Difficulties in Transparent and Opaque Orthographies: A Comparison Between Scandinavian and U.S./Australian Children. *Dyslexia, 16*, 119-142.

***Spelling*** an error detection (i.e., proofreading) task. Participants read a 270-word passage, which contains 30 spelling errors, and were asked to underline the misspelled words. This test has been validated in---and reference scores exist for---University students. The score is the number of misspelled words that were underlined within a time limit of 2 minutes. (Correctly spelled words that are underlined are ignored.)

Reference for the test battery including this spelling test:

Strømsø, H., Hagtvet, B. E., Lyster, S. A. H., & Rygvold, A.-L. (1997). Lese- og skriveprøver for studenter på høyskole- og universitetsnivå [Reading and writing tests for students at college and university level]. Department of Special Needs Education, University of Oslo.

**Russian**

***Author Recognition Test*** Adapted for Russian from Stanovich & West, 1989. In this test the subject is offered a list of 100 names, among which he or she needs to find the names of authors. The names of real authors make up half of this list, the other half includes fillers (real names of people who are not authors). The names of 50 real authors were chosen based on best-selling lists and literary award lists, half of them were Russian-language authors and the others were authors whose works were translated into Russian. No authors from the school curriculum were included.

The final score is calculated by subtracting the number of wrongly selected foils from the number of correctly selected authors (max=50).

<http://webanketa.com/forms/68v34e1q74qk0e9m64t6cd9g/>

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402-433.

***Spelling Test*** developed based on the Spelling recognition test for the English language (Andrews & Hersch, 2010). 44 Russian words that are often misspelled were selected. Half of the words were presented in the correct spelling, and another half with a spelling error. The task was to determine whether the word is spelled correctly or not.

The final score is counted as the number of correct answers (max=44).

<https://webanketa.com/forms/68v38chp6wqkarb36gw3ee9j/>

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, *139*(2), 299.

***Vocabulary test*** Online vocabulary test developed by Grigorii Golovin (2015) (https://www.myvocab.info/). The test was constructed according to Item Response Theory and Computerized Adaptive Testing. It involves a lexical decision task (participants are presented with words of different frequency and pseudowords) with some control questions (to specify the meaning of the word in a multiple choice task).

The final score is the estimated approximate vocabulary size in thousands of words.

The test reliability is 0.95 (Golovin 2015).

Головин Г.В. Измерение пассивного словарного запаса русского языка. Социо- и психолингвистические исследования. 2015. Вып. 3. С. 148–159. [Golovin G.V. Izmerenie passivnogo slovarnogo zapasa russkogo yazyka [Measuring the Passive Vocabulary of the Russian Language]. Sotsio- i psikholingvisticheskie issledovaniya [Social and Psycholinguistic Studies], 2015, vol. 3, pp. 148–159.]: http://splr.psu.ru/wp-content/uploads/2015/12/%D0%93%D0%BE%D0%BB%D0%BE%D0%B2%D0%B8%D0%BD\_2015.pdf).

**Spanish**

***Lexical decision task***The Lextale-Esp is a Spanish version of the Lextale (Izura, Cuetos, & Brysbaert, 2014). The task was administered in a computerized form using Psychopy software (Peirce, Gray, Simpson, MacAskill, Höchenberger, Sogo, Kastman & Lindeløv, 2019). Participants saw a string of letters and they were instructed to press as fast as possible the key ‘s’ if they could recognize the string as a word that belonged to Spanish or ‘n’ if they could not. Task duration was approximately 5 minutes. Raw results .csv files include the correct response (column ‘rta\_correcta’), the key pressed by the participant (column ‘key\_resp.keys’), whether the response was correct or incorrect (column ‘key\_resp.corr’) and the response time for each stimulus (column ‘key\_resp.rt’).

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicologica*, 35, 49-66.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

***Rapid Automatized Naming Task***Participants’ reading skills were operationalized using rapid automatized naming speed, which was measured using a software developed by Bruno Nicenboim (https://github.com/bnicenboim/py-ran-task). The procedure of the test was the following: Each subject was instructed to read a series of trials with 50 items; the items were the same set of letters or numbers that were used in Denckla and Rudel (1976): {a, s, d, p, g} and {2, 6, 9, 4, 7}. The first eight trials are composed of letters and the following eight ones of numbers. The items are displayed in five rows of ten columns and are listed in random order. The test starts with two practice trials to familiarize the participants with the task*.* The software measures average reading time for each set of letters or numbers in seconds. Participants were instructed to start reading aloud as fast as possible immediately after pressing the spacebar, and to press it again immediately after finishing reading aloud the last item. In case they misread, they were instructed to reread only the misread item. Task duration was between 5 and 8 minutes

Denckla, M. B. and Rudel, R. G. (1976), Rapid ’Automatized’ Naming (R.A.N): Dyslexia Differentiated from other learning disabilities, *Neuropsychologia*, 14, 471–479

***Word and non-word decoding*** Rapid reading ability was measured with the word and non-word reading tasks of the PROLEC-SE battery (Ramos Sánchez & Cuetos Vega, 2005)

Word reading was assessed via the Word Reading subtest***.*** Participants were instructed to read aloud a list of 40 individual words, divided into rows of four words. Participants were scored one point for each correct answer. The experimenter recorded the reading session and reading speed was calculated from the moment the participant started reading until she finished. The list of words includes 20 short and 20 long words, counterbalanced according to their frequency, leading to 10 short very frequent words (FC), 10 short infrequent words (IC), 10 long very frequent words (FL) and 10 long infrequent words (IL). Test duration was between 1 and 2 minutes.Non-word reading was assessed via the Non Word subtest. Participants were instructed to read aloud 40 individual pseudo-words (i.e., non-words adhering to the phonotactic constraints of Spanish) divided into rows of four pseudo-words. Participants were scored one point for each correct answer. The experimenter recorded the reading session and reading speed was calculated from the moment the participant started reading until she finished. The list of non-words includes 20 short and 20 long non-words, counterbalanced according to their syllable complexity, leading to 10 short non-words with simple syllables (SC), 10 short non-words with complex syllables (CC), 10 long non-words with simple syllables (SL), 10 long non-words with complex syllables (CL). Short non-words comprise two syllables, and long words comprise four syllables. Simple syllables comprise consonant-vowel (CV) graphemes and complex syllables are composed by CCV, CCVC and CVVC graphemes. Test duration was between 1 and 2 minutes.

Ramos Sánchez, J. L. & Cuetos Vega, F. (2005). *PROLEC-SE, Evaluación de los procesos lectores en alumnos del Tercer ciclo de Educación Primaria y educación Secundaria Obligatoria,* *3ra* *Edición* *revisada*. Madrid:TEA Ediciones.

**Turkish**

***KOBIT*** (Babur et al., 2011) was used as the offline L1 test for Turkish experiments (stands for Kelime Okuma Bilgisi Testi, translation “Word Reading Knowledge Test). It is similar to TOWRE, first Turkish real words are read aloud for 1 minutes, followed by pseudowords for 1 minutes. The results were reported as the number of words that were read correctly, without stammering.

Babür, N., Haznedar, B., Erçetin, G., Özerman, D., & Çekelek, E. E. (2011). Türkçe'de Kelime Okuma Bilgisi Testi'nin (KOBİT) Geliştirilmesi. Boğaziçi Üniversitesi Eğitim Dergisi, 28(2), 1-21.